Winning Space Race with Data Science

Daria Efimova
November 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- To predict the price of each launch, I'm using public information about a Space X rocket launches via SpaceX API and web scrapping from Wikipedia website. I analysed collected data with exploratory data analysis and tried to determine if SpaceX will reuse the first stage by training different ML models.
- I found out that the number of flight, the payload mass and the orbit are important factors which affect successful landing of the first stage. Also we can use either Logistic Regression, SVM or KNN to predict first stage successful landing.

Introduction

- Our company has a plan to launch rockets and making it with a relatively inexpensive cost. The SpaceX company is already doing it by reusing the first stage of the rocket and we would like to compete with them.

- So, we also need the first stage landing successfully. To be sure it’s happening, we would like to analyse the SpaceX public information to predict if their first stage would land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Webscrapping from the table with info about launches for Wikipedia website
- Perform data wrangling
  - I used Exploratory Data Analysis to find some patterns in the data and determine what would be the outcome for training ML models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - To predict the landing successful rate I was choosing the best model between Logistic Regression, SVM, KNN and Decision Tree and tuned it's hyperparameter using GridSearchCV.

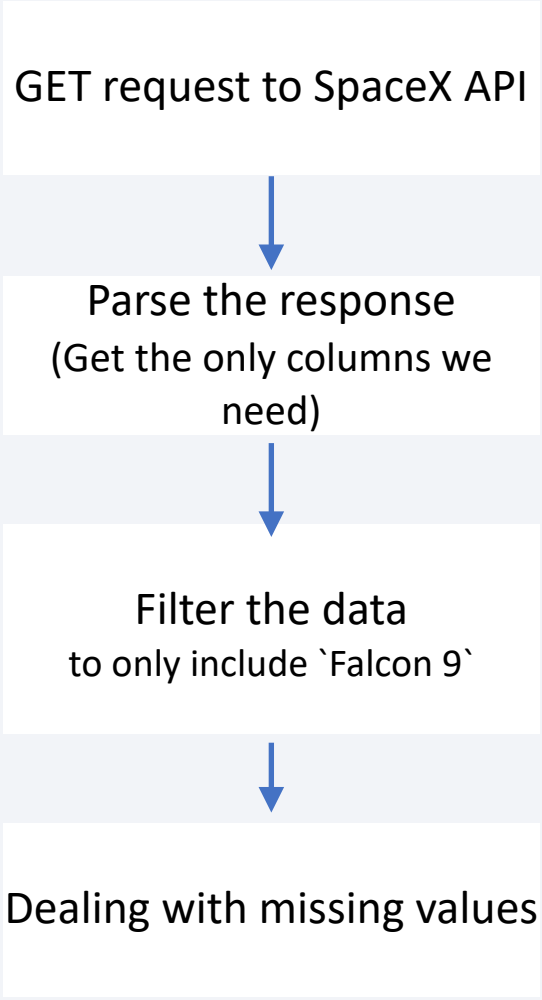
Data Collection

Data sets were collected from SpaceX API and Wikipedia website using webscrapping

Data Collection – SpaceX API

- First I did a GET request to API, then got the data I needed from the response and made some data preparation, like filter Booster Version column and replace missing PayloadMass values with it's mean

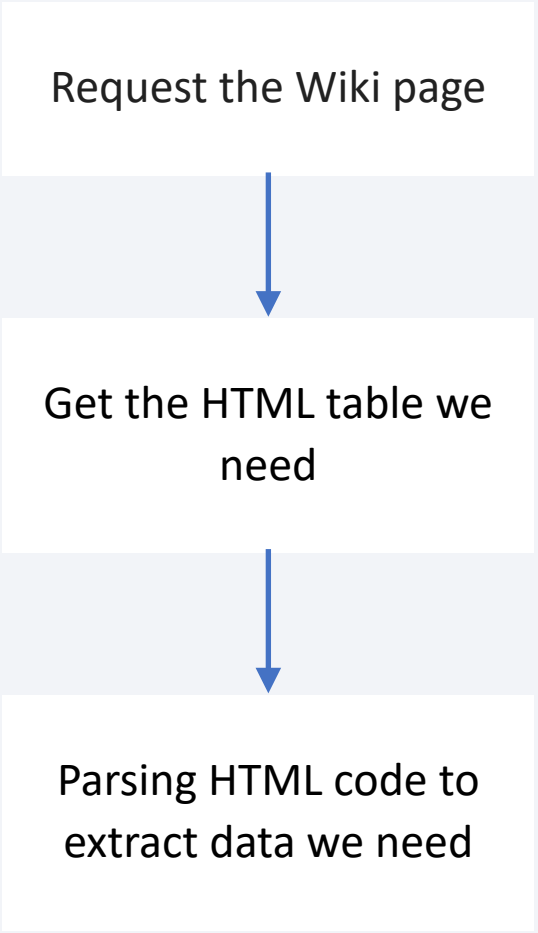
- https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

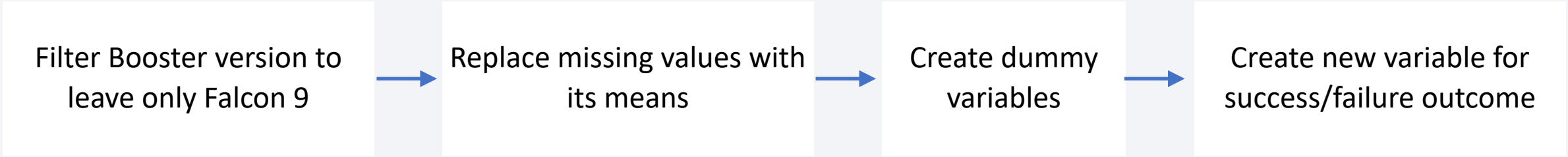
- Request the Falcon9 Launch Wiki page from its URL, then extract the HTML tables which have data we need, get all column/variable names from header of that table and create a data frame by parsing the launch HTML tables.

- https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/jupyter-labs-webscraping.ipynb



Data Wrangling

- First I filtered Booster Version to keep only ‘Falcon 9’ launches and replace missing PayloadMass values with its mean. I also created dummy variables for categorical features. Then I take a look at what the outcomes dataset has and see 4 different values represent a failure to land, so I created a new column ‘Class’ which has only 2 value: success of failure to land of first stage of the rocket.
- https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- This charts were plotted:
  - Flight Number vs Launch Site.
  - Payload vs Launch Site.
  - Success rate of each orbit type
  - FlightNumber vs Orbit type.
  - Payload vs Orbit type.
  - The launch success yearly trend

- https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The SQL queries I performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string ‘CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass.
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- I added to a folium map this objects:
  - Circles and markers of all launch sites to see where this launch sites are located on a map.
  - For each site I added a marker cluster to mark the success/failed launches on this site.
  - Also I added MousePosition to get coordinates of the objects close to launch sites I would like to explore (coastline and city of LA) and lines to this objects.
- https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- I made a SpaceX Launch Records Dashboard with pie chart of Total Success Launches for each Launch site and scatterplot for Correlation between Payload and Success for each Launch Site with Payload slider.
- https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/Dashboard.ipynb

Predictive Analysis (Classification)

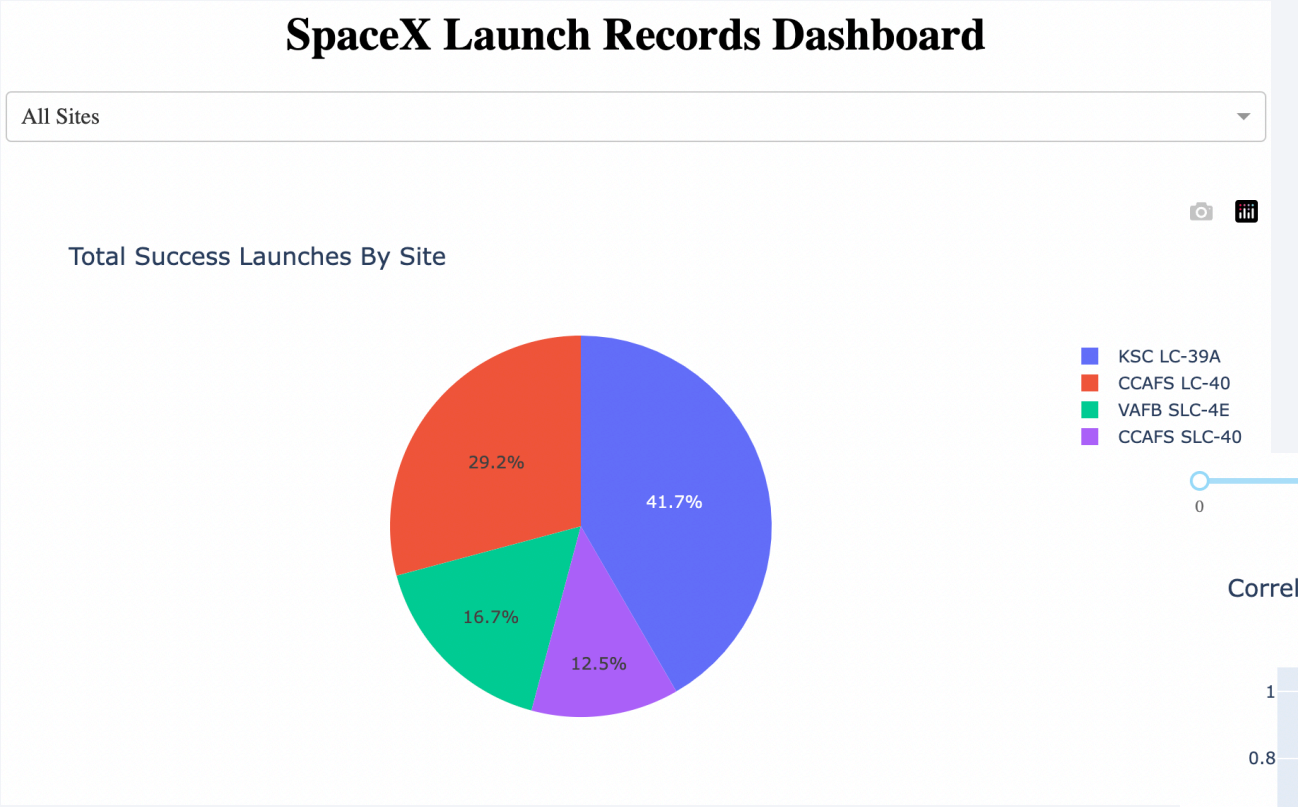
- To predict if first stage will land successfully, I built SVM, Classification Trees, Logistic Regression and KNN ML models and choose the best of it by calculating score. SVM, Classification Trees and Logistic Regression has hyperparameters and its best values need to be choose, I did it with GridSearchCV.
- https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

As a result of exploratory data analysis I could make this conclusions:

- Different launch sites have different success rates. CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- On the VAFB-SLC launch site there are no rockets launched for heavy payload mass.
- As lower orbit is, as better success rate it has.
- Also in the LEO orbit the success appears related to the number of flights.

Results


SpaceX Launch Records Dashboard
All Sites
Total Success Launches By Site
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40


0
2500
5000
7500
10000
Correlation between Payload and Success for all Sites and 0
class
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
10k
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

Results

I’ve done a predictive analysis and find out that the most accurate model to predict the chance of first stage successful landing is Decision Tree. It gives us a correct result in 86% of cases.

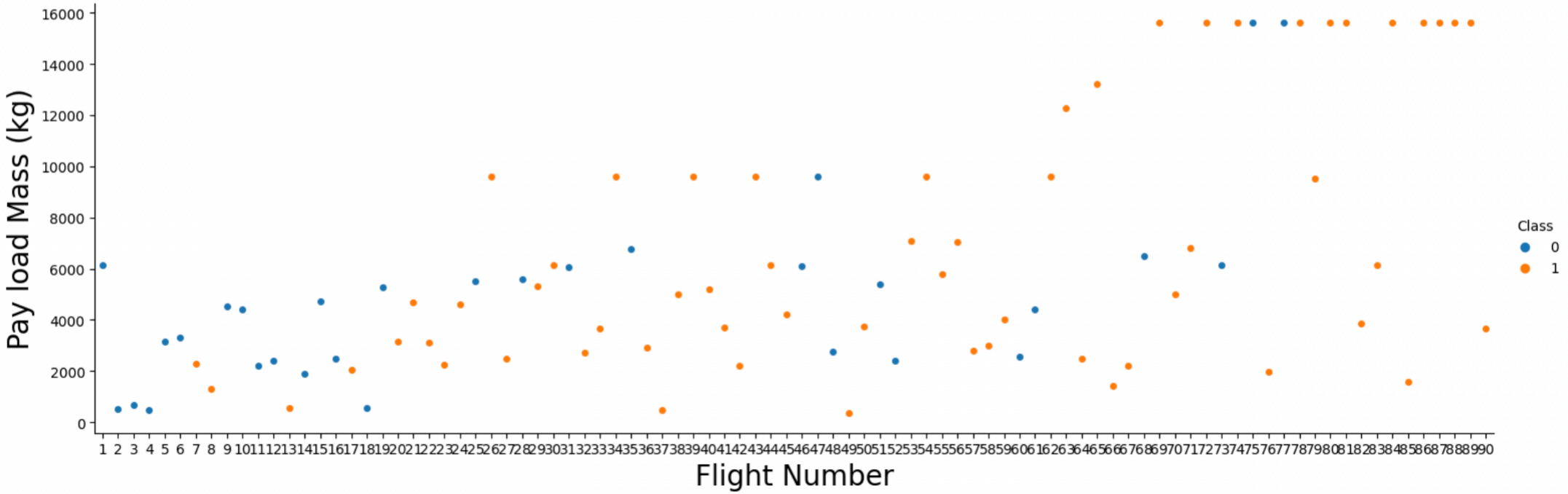
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

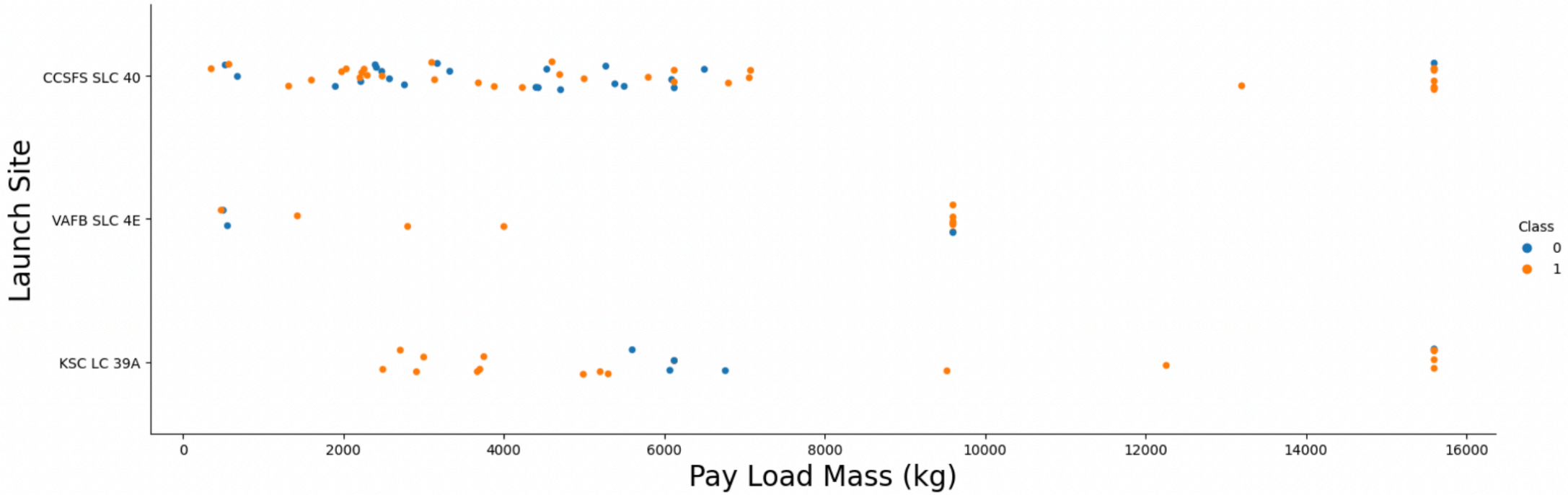
We can plot out the `FlightNumber` vs. `PayloadMass` and overlay the outcome of the launch. We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

[21]:
```
sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 3)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Pay load Mass (kg)",fontsize=20)
plt.show()
```



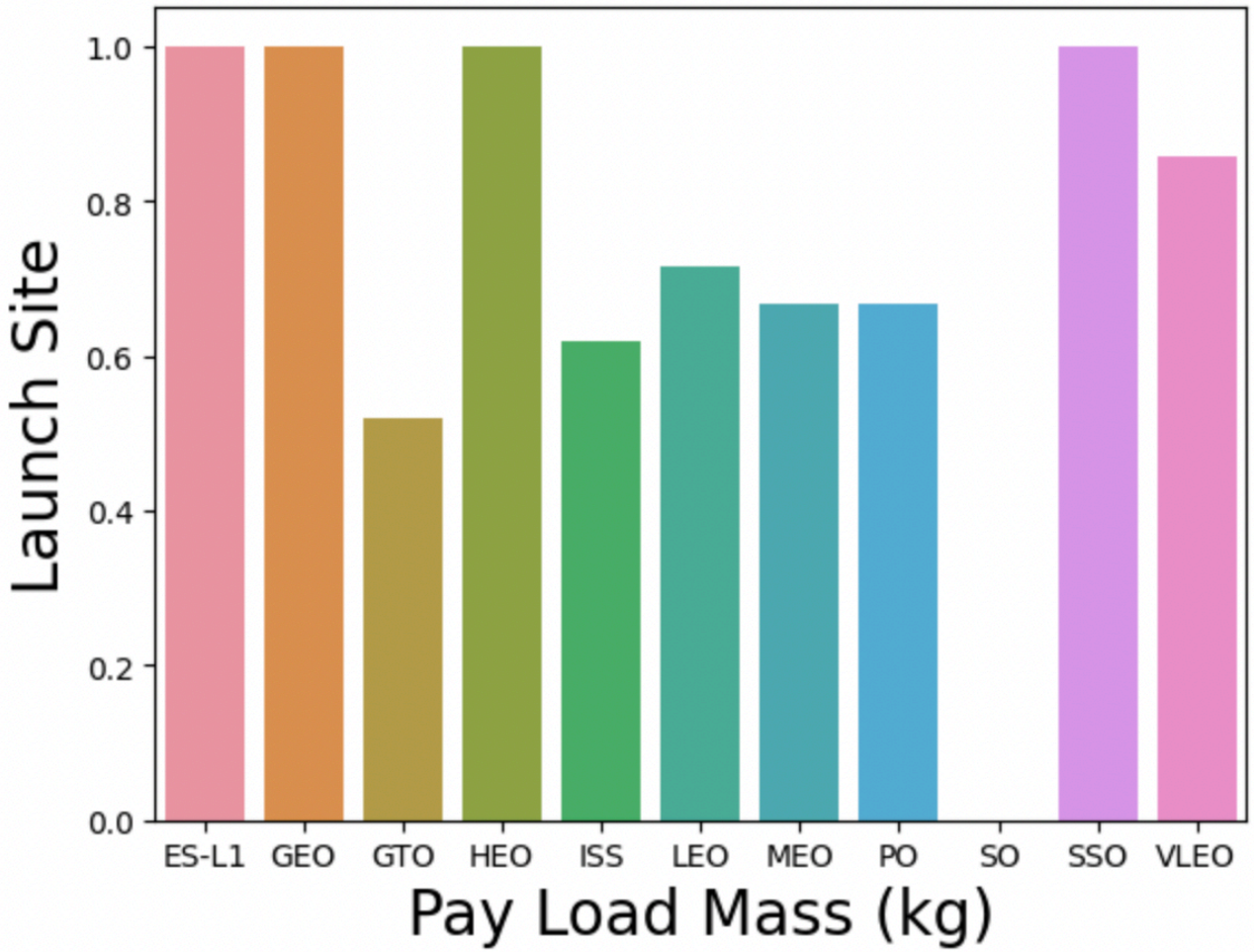
Payload vs. Launch Site

[22]:
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 3)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



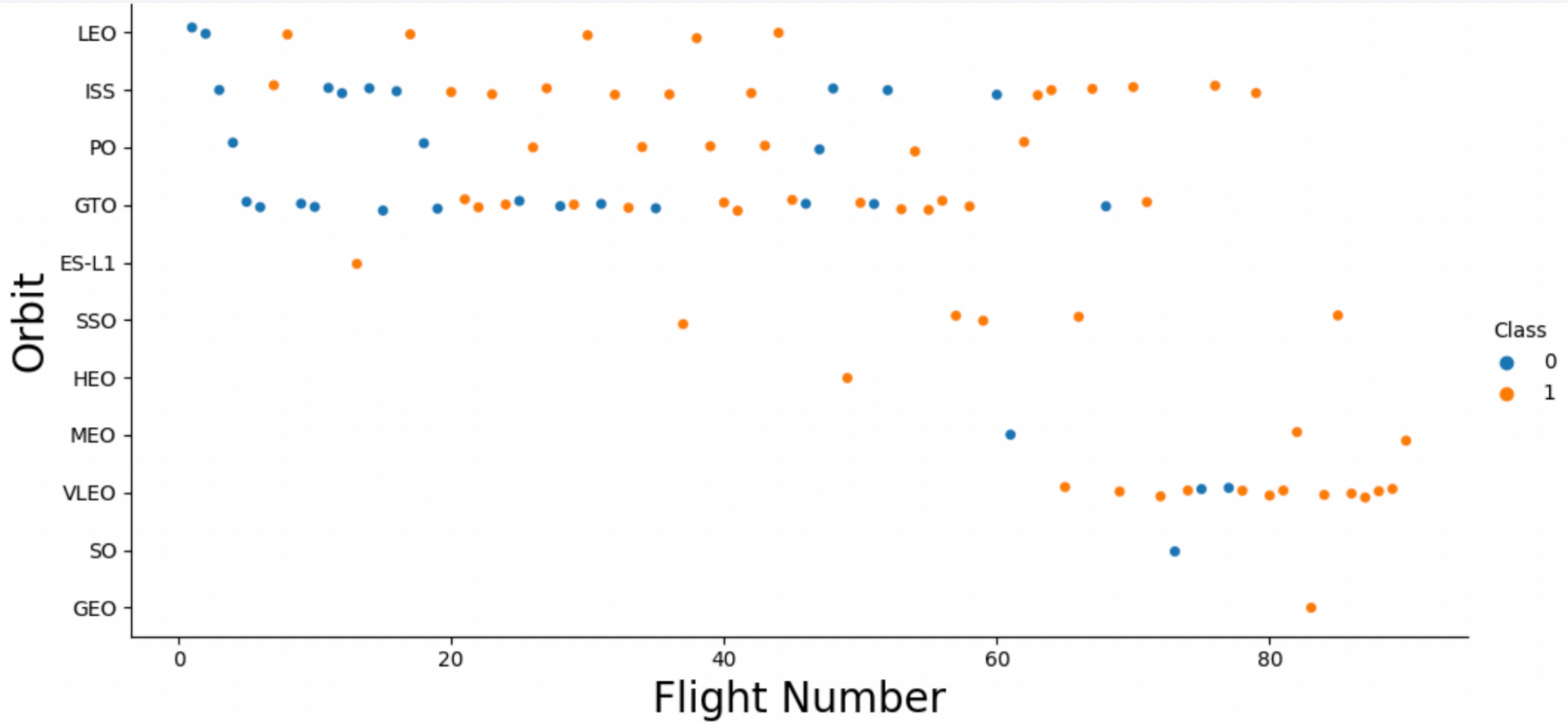
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



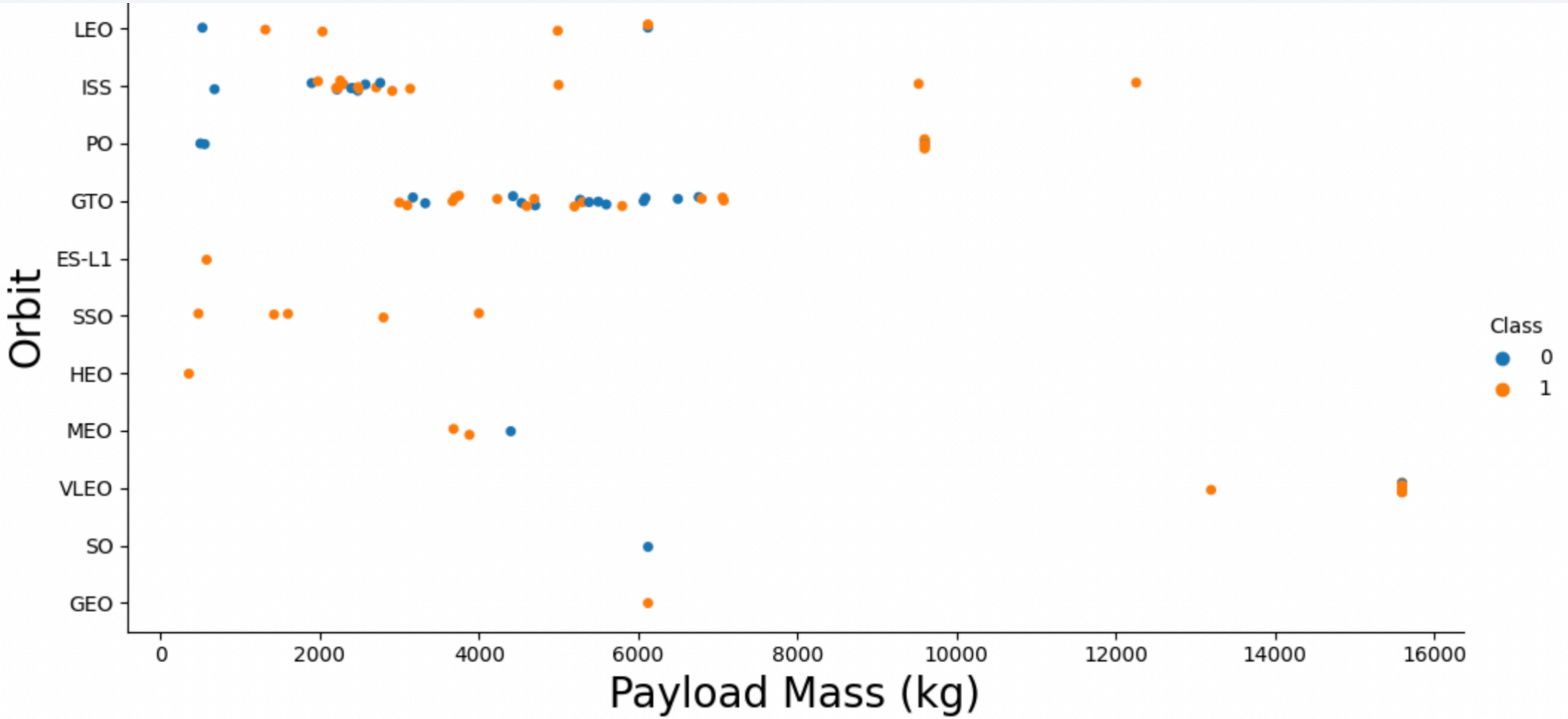
We can see that as lower orbit is, as better success rate it has.

Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

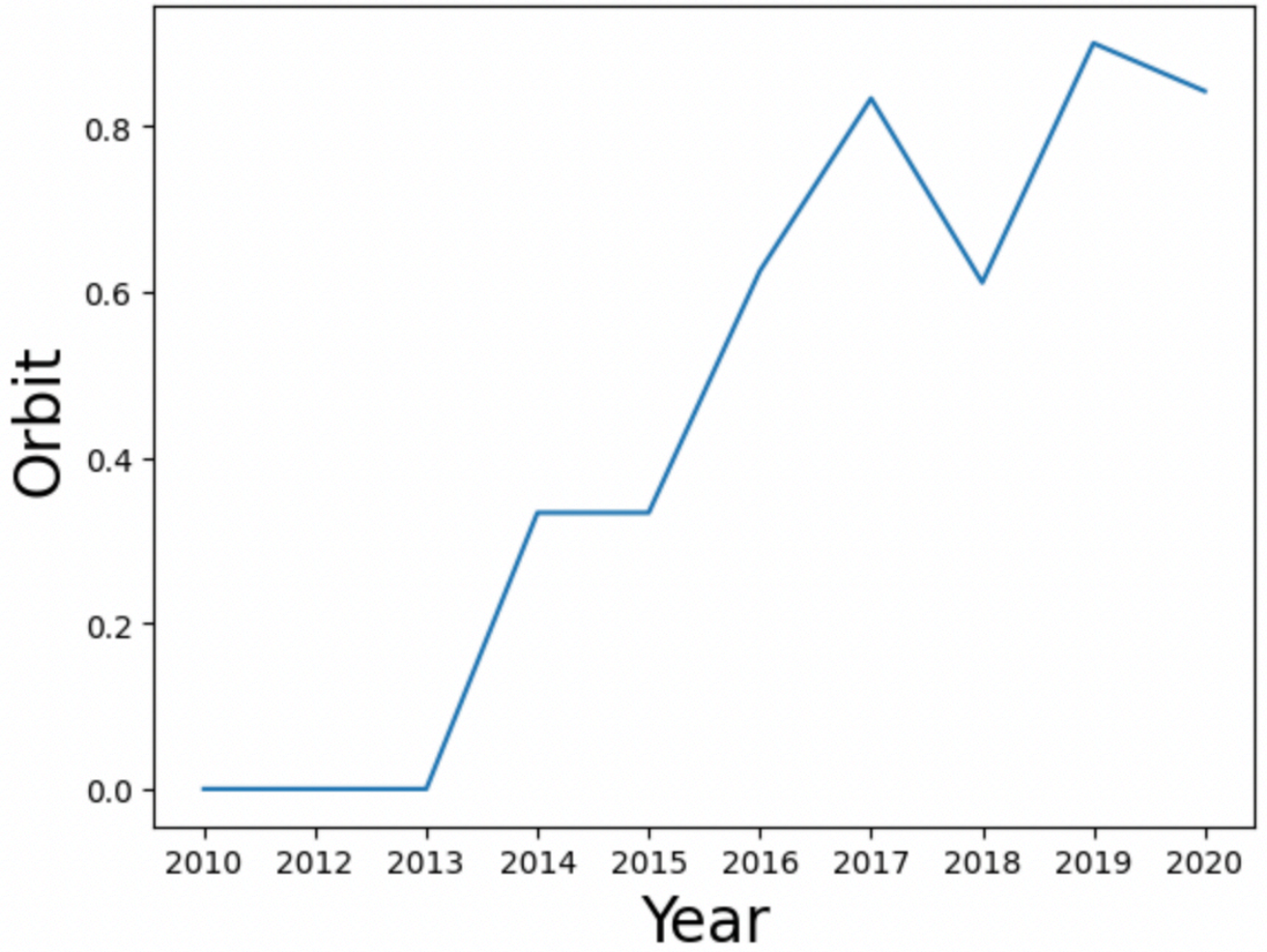
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
cursor.execute('SELECT DISTINCT Launch_Site FROM Spacex')
cursor.fetchall()
```

```
[('CCAFS LC-40',), ('VAFB SLC-4E',), ('KSC LC-39A',), ('CCAFS SLC-40',)]
```

As you can see, the Launch sites names are: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
sql_string = "SELECT * FROM Spacex WHERE Launch_Site LIKE 'CCA%' LIMIT 5"
df = pd.read_sql(sql_string, conn)
df
```

	index	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
0	0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
1	1	08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success
2	2	22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
3	3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
4	4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
cursor.execute("SELECT sum(PAYLOAD_MASS__KG_) FROM Spacex WHERE Customer LIKE 'NASA (CRS)%'")
cursor.fetchall()
```

```
[(48213,)]
```

As you can see, the total payload mass launched by NASA is 48213 kg.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
cursor.execute("SELECT avg(PAYLOAD_MASS__KG_) FROM Spacex WHERE Booster_Version LIKE 'F9 v1.1%'")
cursor.fetchall()
```

```
[(2534.6666666666665,)]
```

As you can see, the average payload mass carried by booster version F9 v1.1 is 2534.67 kg.

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
sql_string = "SELECT min(Date) FROM Spacex WHERE [Landing _Outcome] LIKE 'Success (ground pad)'"
cursor.execute(sql_string)
cursor.fetchall()
```

```
[('01-05-2017',)]
```

As you can see, the first successful landing outcome in ground pad was achieved on 1 May 2017

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
sql_string = "SELECT DISTINCT Booster_Version FROM Spacex WHERE [Landing _Outcome] LIKE 'Success (drone ship)' "+
"AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000"

cursor.execute(sql_string)
cursor.fetchall()
```

```
[('F9 FT B1022',), ('F9 FT B1026',), ('F9 FT  B1021.2',), ('F9 FT  B1031.2',)]
```

As you can see, the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
sql_string ="SELECT (SELECT count([Landing _Outcome]) FROM Spacex WHERE [Landing _Outcome] LIKE 'Success%') "+
"as Success, (SELECT count([Landing _Outcome]) FROM Spacex WHERE [Landing _Outcome] NOT LIKE 'Success%') "+
"as Failure FROM Spacex"
cursor.execute(sql_string)
cursor.fetchone()
```

```
(61, 40)
```

As you can see, mission has 61 of successful landings of first stage and 40 were ended with failure.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
sql_string = "SELECT DISTINCT Booster_Version FROM Spacex WHERE PAYLOAD_MASS__KG_ ="+
" (SELECT max(PAYLOAD_MASS__KG_) FROM Spacex)"

cursor.execute(sql_string)
cursor.fetchall()
```

```
[('F9 B5 B1048.4',),
 ('F9 B5 B1049.4',),
 ('F9 B5 B1051.3',),
 ('F9 B5 B1056.4',),
 ('F9 B5 B1048.5',),
 ('F9 B5 B1051.4',),
 ('F9 B5 B1049.5',),
 ('F9 B5 B1060.2 ',),
 ('F9 B5 B1058.3 ',),
 ('F9 B5 B1051.6',),
 ('F9 B5 B1060.3',),
 ('F9 B5 B1049.7 ',)]
```

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
sql_string = "SELECT DISTINCT Booster_Version, Launch_Site FROM Spacex WHERE [Landing _Outcome] "+
"LIKE 'Failure (drone ship)%' AND Date LIKE '%2015'"
cursor.execute(sql_string)
cursor.fetchall()
```

```
[('F9 v1.1 B1012', 'CCAFS LC-40'), ('F9 v1.1 B1015', 'CCAFS LC-40')]
```

As you can see, in year 2015 only the boosters F9 v1.1 B1012 and F9 v1.1 B1015 were launched from the CCAFS LC-40 site

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

```
sql_string = "SELECT [Landing _Outcome], COUNT([Landing _Outcome]) "+
"FROM Spacex WHERE Date BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY [Landing _Outcome] "+
"ORDER BY [Landing _Outcome] DESC"

cursor.execute(sql_string)

cursor.fetchall()
```

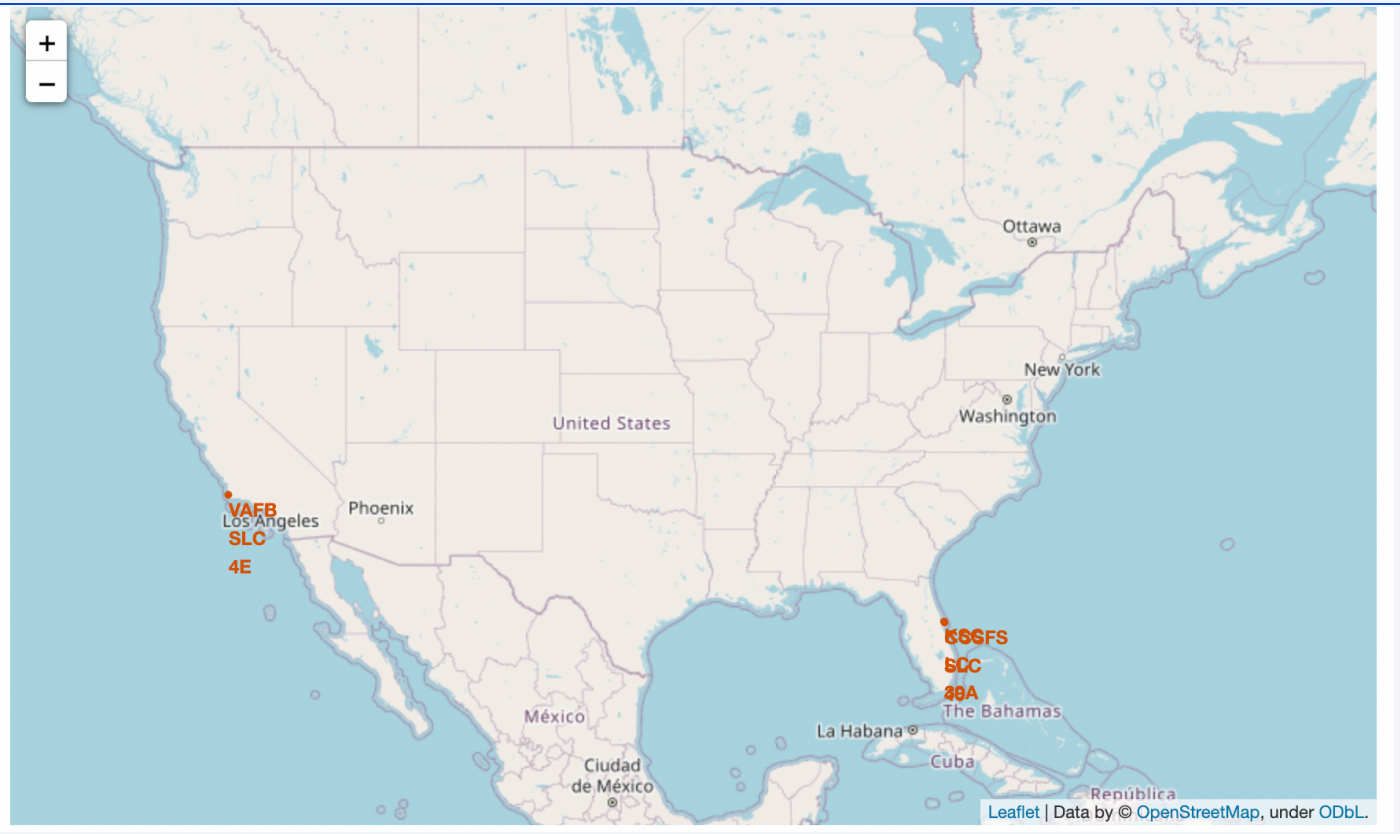
```
[('Success (ground pad)', 6),
 ('Success (drone ship)', 8),
 ('Success', 20),
 ('No attempt ', 1),
 ('No attempt', 10),
 ('Failure (parachute)', 2),
 ('Failure (drone ship)', 4),
 ('Failure', 3),
 ('Controlled (ocean)', 3)]
```

You can see that between 6 June 2010 and 20 March 2017 were launched a lot of rockets which ended with different result in terms of landing it's first stage.

Section 3

Launch Sites Proximities Analysis

All launch sites on map



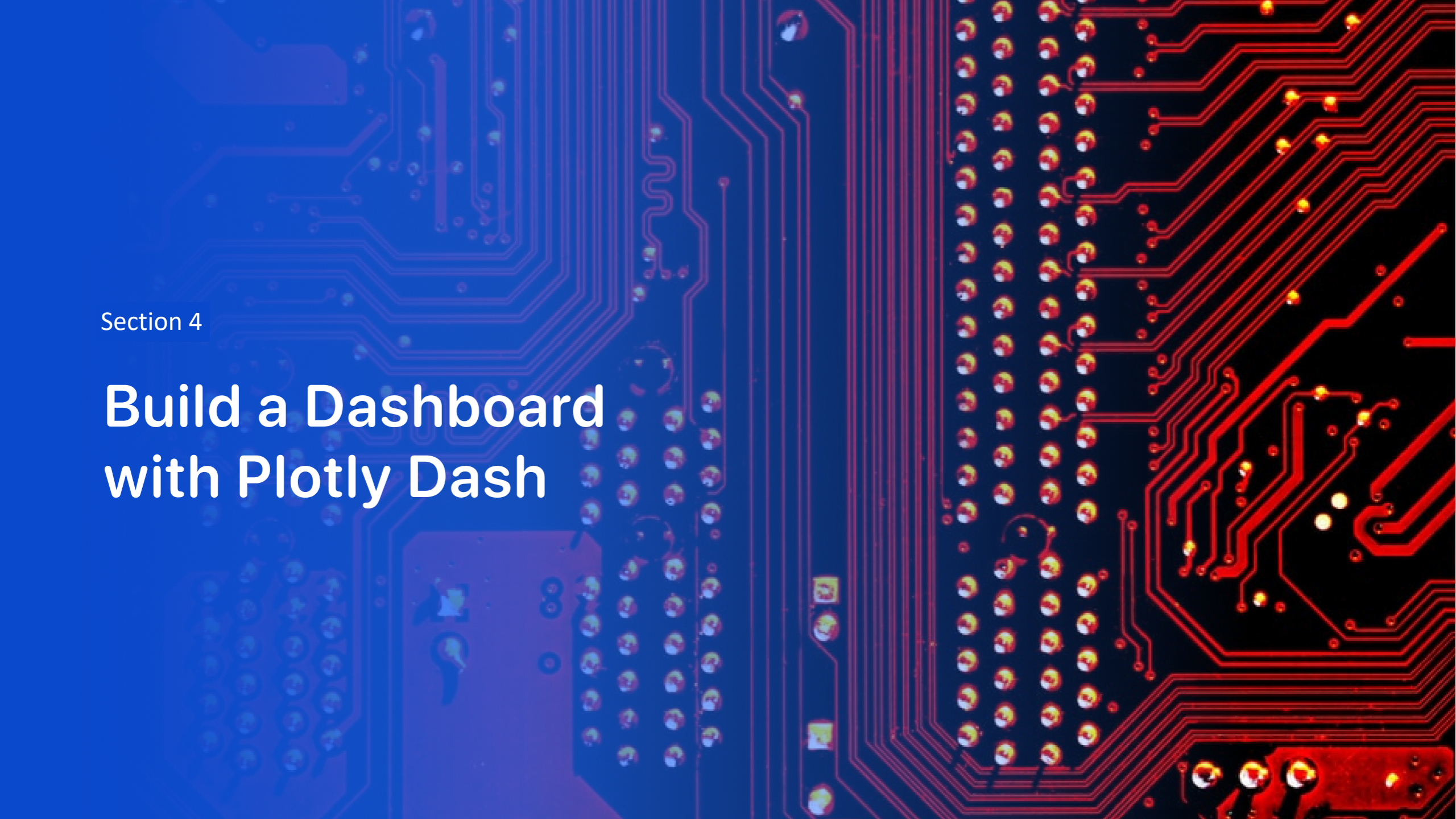
As you can see on the map, all launch sites located in USA, on the different coasts, as close to equator as it could be in USA.

Launch outcomes on the map

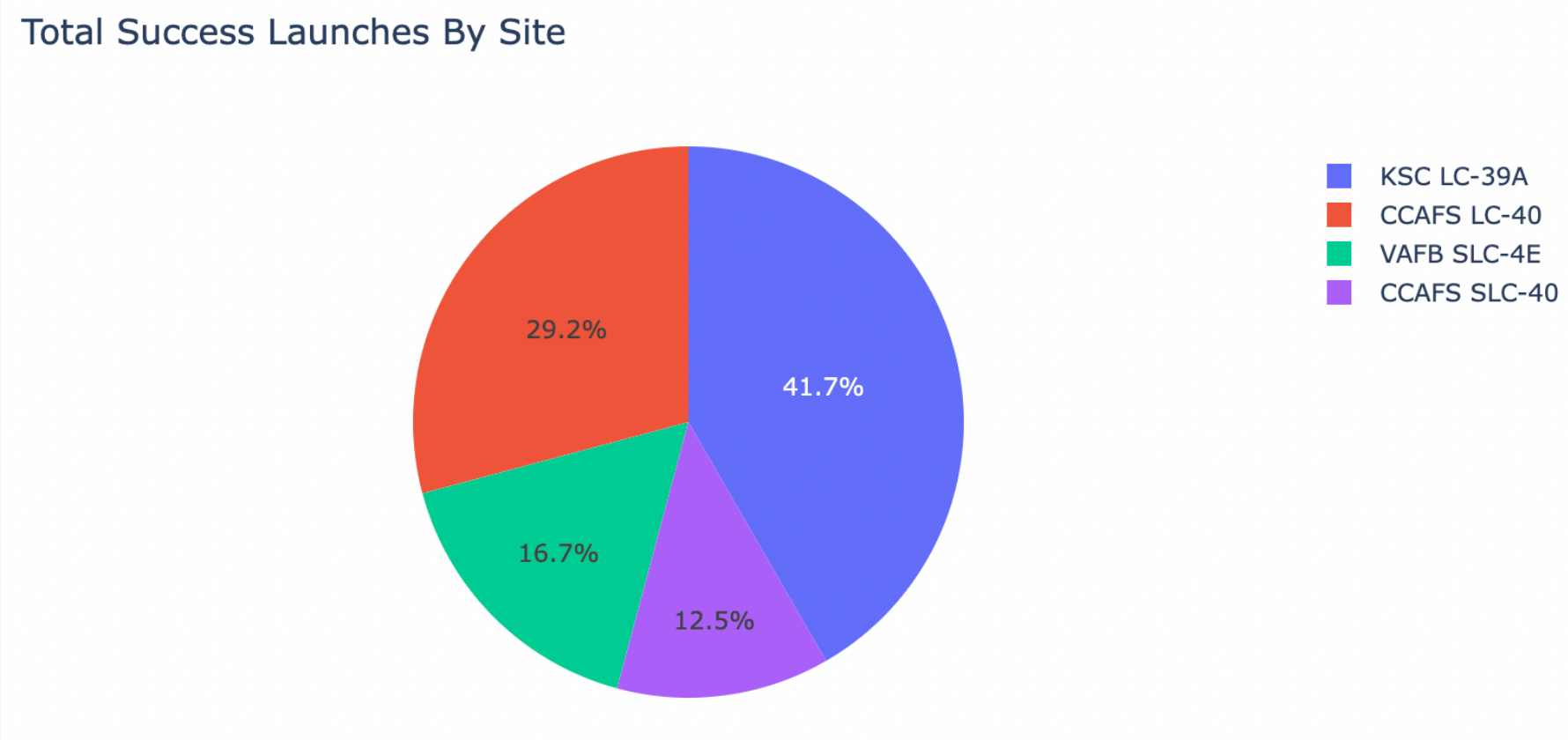
As you can see on the map, launch site called ‘VAFB SLC-4E’ has 13 launches, in which 3 were failed landing the first stage and 10 were successful. Another 77 launches were handled by another launch sites.

Distances between a launch site and objects near by

On the map we can see that launch site called ‘VAFB SLC-4E’ is located in 1.38km away of the coast and around 240km away of the nearest big city (which is Los Angeles)

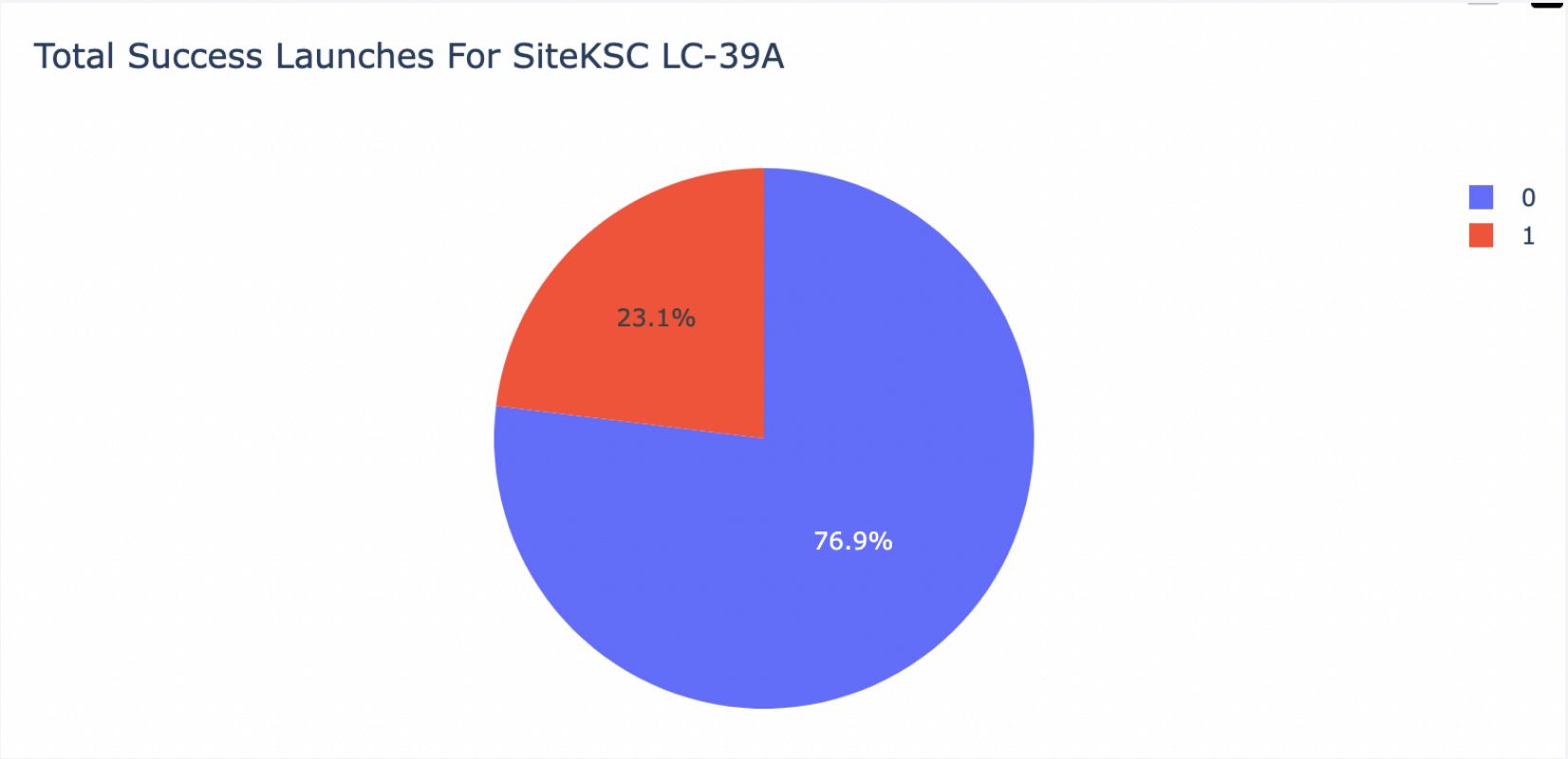

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches By Site



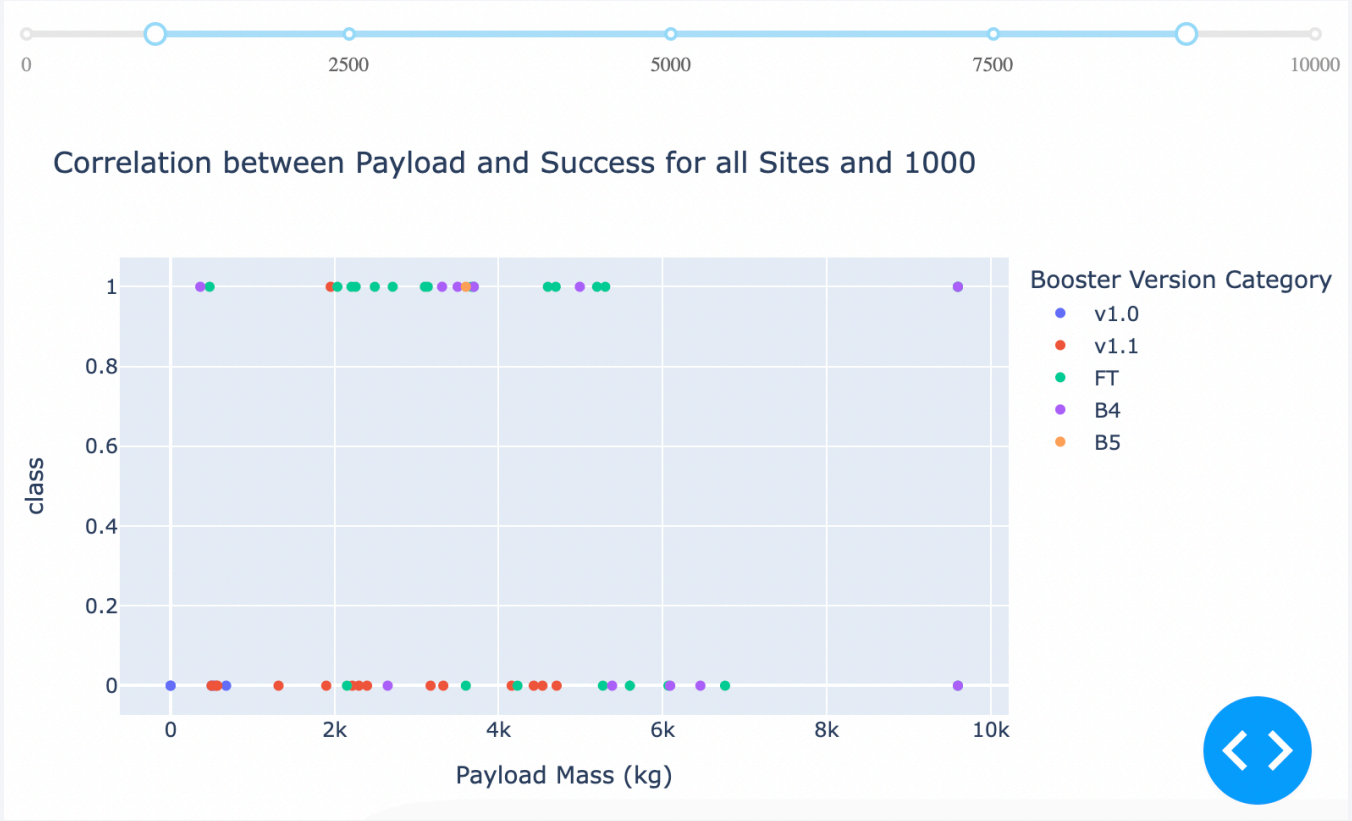
As you can see, the most successful site is ‘KSC LC-39A’ and the least successful is ‘CCAFS SLC-40’

Total Success Launches For The Site With The Highest Success Ratio



As you can see, there are almost 77% of the launches ended with the landing first stage successfully and only 23% of the launches are failed.

Payload vs. Launch Outcome scatter plot for all sites



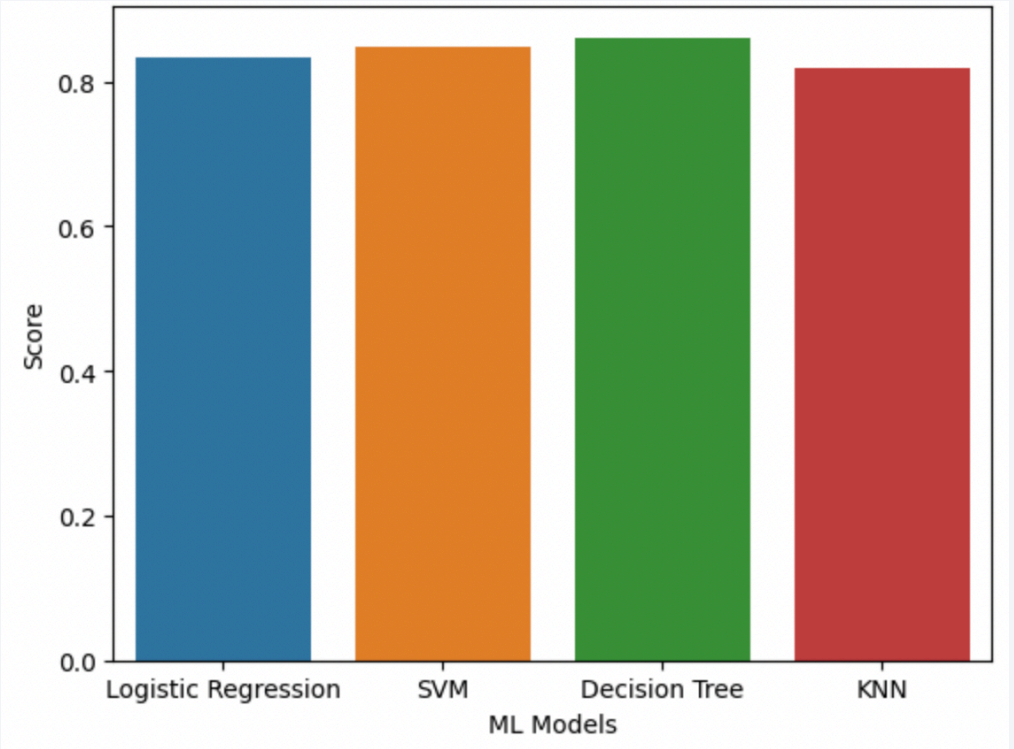
As you can see, the payload range from around 2000kg to 4000kg has the largest success rate, same as it looks like the FT Booster also the most successful one.

Section 5

Predictive Analysis (Classification)

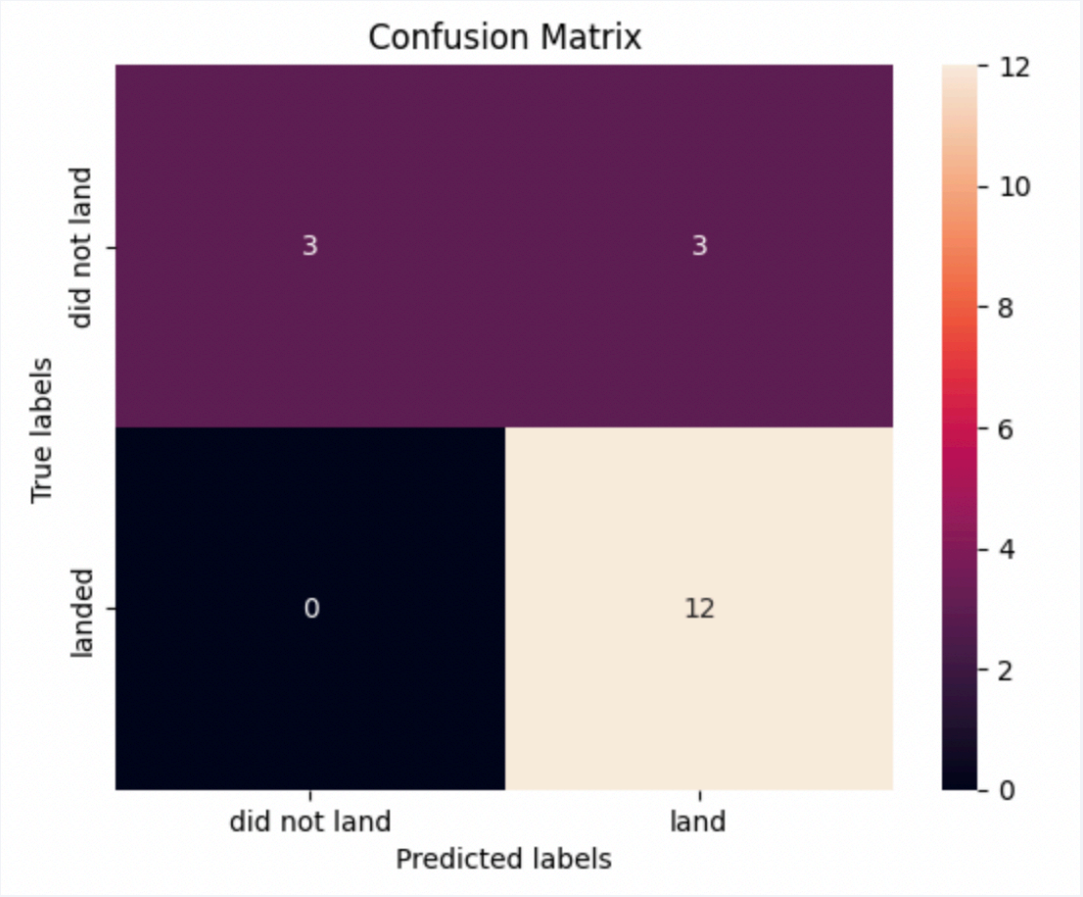
Classification Accuracy

You can see here that Decision Tree model has the highest classification accuracy.



Confusion Matrix

You can see here the confusion matrix for decision tree model. It shows that decision tree can distinguish between the different classes. Also you can see that the major problem is false positives.



Conclusions

- The most accurate model to predict the chance of first stage successfully landing is decision tree. It gives us a correct result in 86% of cases.
- Different launch sites have different success rates. CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- On the VAFB-SLC launch site there are no rockets launched for heavy payload mass.
- As lower orbit is, as better success rate it has.
- Also in the LEO orbit the success appears related to the number of flights.

Appendix

- Datasets which I created during this project:
  - https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/dataset_part_1.csv
  - https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/dataset_part_2.csv
  - https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/dataset_part_3.csv
  - https://github.com/datascientist211/SpaceXPrediction/blob/6ea5cea6d8747e41f48baaf4dabcaec09626f3f5/spacex_web_scraped.csv

Thank you!